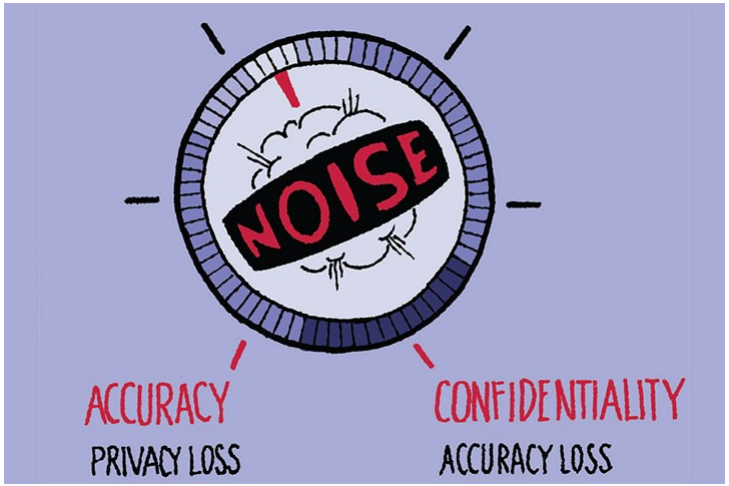
Lecture 29

Private Data Analysis



Announcements

- Course project deliverable 2 due Friday @ 11:59pm

- Office hours every Tuesday (noon to 4pm) and Friday (11am to 1pm)
  - Link to book appointments is on the course website

Today

Notions of privacy

Framework called differential privacy

Setting the Stage

Data drives modern decision-making

- Healthcare, public policy, economics, business

Sensitive data

- People's data in these domains often contain sensitive information



Setting the Stage

Data drives modern decision-making

- Healthcare, public policy, economics, business

Sensitive data

- People's data in these domains often contain sensitive information

Challenge

- How can we reliably learn interesting things from data without compromising individual privacy?



Private Data Analysis

Imagine we have a dataset containing one record for each person and we perform some data analysis

Informally: an analysis is **private** if it does not leak very much of any person's information



Private Data Analysis

Imagine we have a dataset containing one record for each person and we perform some data analysis



Informally: an analysis is **private** if it does not leak very much of any person's information

Goal: perform useful data analysis while providing meaningful privacy guarantees

- A balance between privacy and utility

Analyzing Student Trends

Suppose the college tasks us with studying our students across some sensitive topics

- How often do students cheat on coursework?
- How many students have a medical emergency each semester?
- How many students regularly use recreational drugs?
- … and so on

Our Study

Goal: estimate percent of students who cheated frequently this semester

- **Utility**: how good is our estimate? (unbiased?, low variance?, etc.)
- **Privacy**: is student information being leaked?

Let's consider some ways to design our study and how they affect privacy and utility

Design 1

Idea: public collection

- Gather a sample of students in a room
- Ask them to raise their hand if they cheated and tabulate the counts

Design 1

Idea: public collection

- Gather a sample of students in a room
- Ask them to raise their hand if they cheated and tabulate the counts

Low utility

- Students are unlikely to answer honestly (biased)

Design 1

Idea: public collection

- Gather a sample of students in a room
- Ask them to raise their hand if they cheated and tabulate the counts

Low utility

- Students are unlikely to answer honestly (biased)

Low privacy

- Others can see your answers just by looking around

Design 2

Idea: private collection

- Privately ask students one-by-one if they cheated and record their responses

Design 2

Idea: private collection

- Privately ask students one-by-one if they cheated and record their responses

Low utility

- Students are still unlikely to answer honestly (biased)

Design 2

Idea: private collection

- Privately ask students one-by-one if they cheated and record their responses

Low utility

- Students are still unlikely to answer honestly (biased)

Medium privacy

- Trust the data collector

Design 2

Idea: private collection

- Privately ask students one-by-one if they cheated and record their responses

Low utility

- Students are still unlikely to answer honestly (biased)

Medium privacy

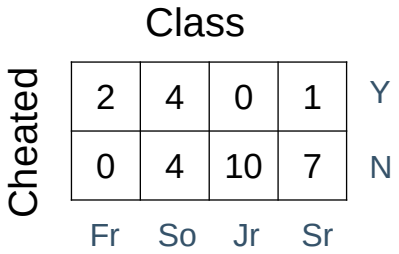
- Trust the data collector

Likely better privacy and utility than Design 1!

Design 3

Idea: anonymous collection

- Students fill out an anonymous Google form
- We also ask for their class (Fr, So, Jr, Sr)



Design 3

Idea: anonymous collection

- Students fill out an anonymous Google form
- We also ask for their class (Fr, So, Jr, Sr)

Cheated	Fr	So	Jr	Sr
Y	2	4	0	1
N	0	4	10	7

Class

Medium utility

- Students are somewhat likely to answer honestly (biased)

Design 3

Idea: anonymous collection

- Students fill out an anonymous Google form
- We also ask for their class (Fr, So, Jr, Sr)

Cheated	Class: Fr	So	Jr	Sr
Y	2	4	0	1
N	0	4	10	7

Medium utility

- Students are somewhat likely to answer honestly (biased)

Medium privacy

- **Mosaic effect**: if someone has outside information that a particular student is a Fr, then they can determine whether that student cheated or not

Design 4

Idea: anonymous collection + randomization

- Students fill out an anonymous Google form
- We also ask for their class (Fr, So, Jr, Sr)
- Add noise to the tabulated counts

Cheated	Class: Fr	So	Jr	Sr
Y	1.26	5.2	0.9	0.2
N	-0.1	4.5	8.42	6.3

High(er) utility

- Students are likely to answer honestly (approx. unbiased)
- Need to round and normalize to get valid counts

High(er) privacy

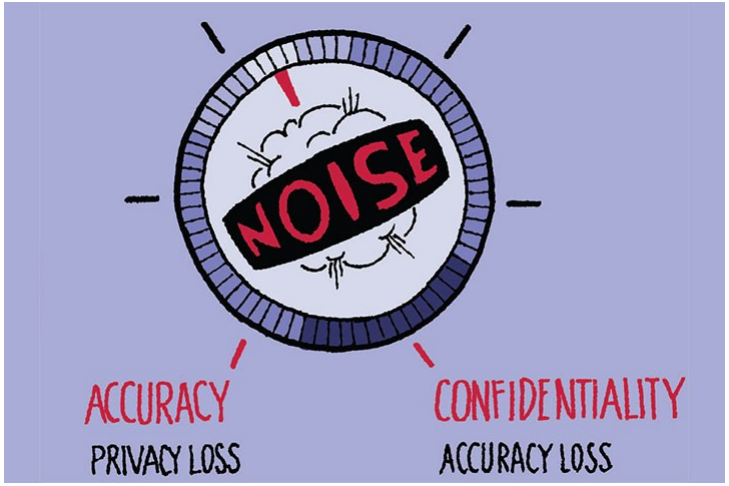
- Noise gives you plausible deniability

Adding Noise

Obtain privacy guarantees using framework of **differential privacy**

Idea: inject randomization (often through adding noise) to the data analysis

- By carefully controlling the randomization, we can provide a guarantee on how much information is leaked in the worst case
- More noise = stronger privacy guarantee



Next Time

How to privatize statistics?

- Mean, median, covariance, etc.
- Not trivial problems!

More details on differential privacy framework